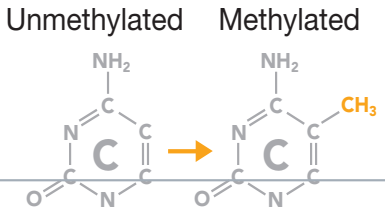
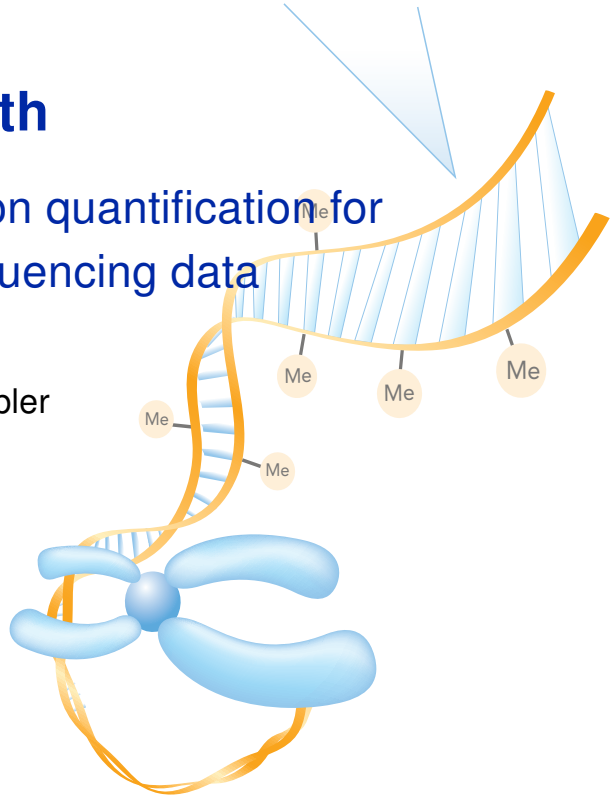
University of Zurich[UZH]
Insitute of Molecular Life Sciences
Institute of Social and Preventive Medicine




# BayMeth

## Improved DNA methylation quantification for affinity capture sequencing data

Andrea Riebler



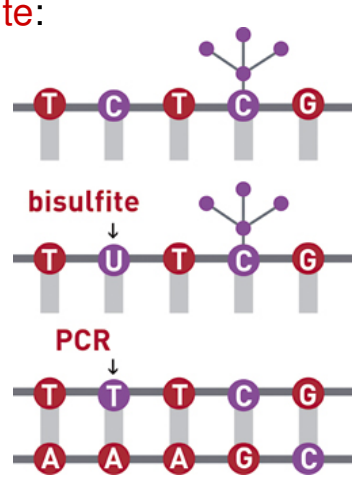

Bioconductor European Developers' Workshop 2012

http://www.illumina.com/Documents/products/datasheets/datasheet_veracode_methylation.pdf

## Technologies for DNA methylation profiling

1. Treatment of DNA with sodium bisulphite:

$$C \to T$$
$$C^M \to C.$$



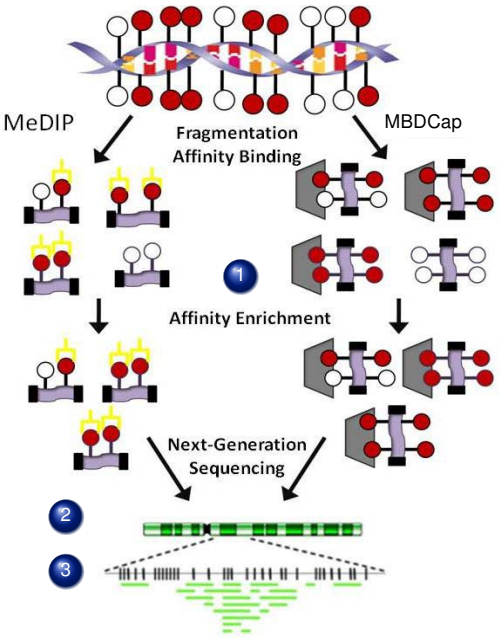

www.diagenode.com/en/applications/bisulfite-conversion.php

2. High-throughput sequencing.

- Advantage: Single base resolution.
- Disadvantage: Whole genome-wide bisulphite sequencing (WGBS) is very expensive and inefficient.
- Methylation arrays are an alternative, but provide less coverage and are only available for human.

## Affinity-capture-based approaches

strike good balance between high cost of WGBS and the low coverage of methylation arrays.




The number of reads mapping to 100bp bins, say, is counted.

$\Rightarrow$ DISCRETE DATA

- Read density not directly interpretable.
- Dependence on CpG density.
- Methods for microarrays not applicable.

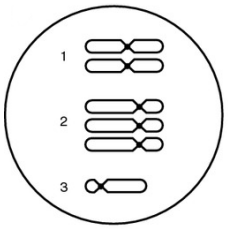


## Statistical analysis: Available software packages

| Software | Reference | Implementation |
|---|---|---|
| `Batman` | Down et al. (Nat Biotech, 2008) | `Java` |
| `MEDIPS` | Chavez et al. (Genome Res, 2010) | `R` / Bioconductor |
| `BALM` | Lan et al. (PLoS ONE, 2011) | `C++` |

A new method is desired that

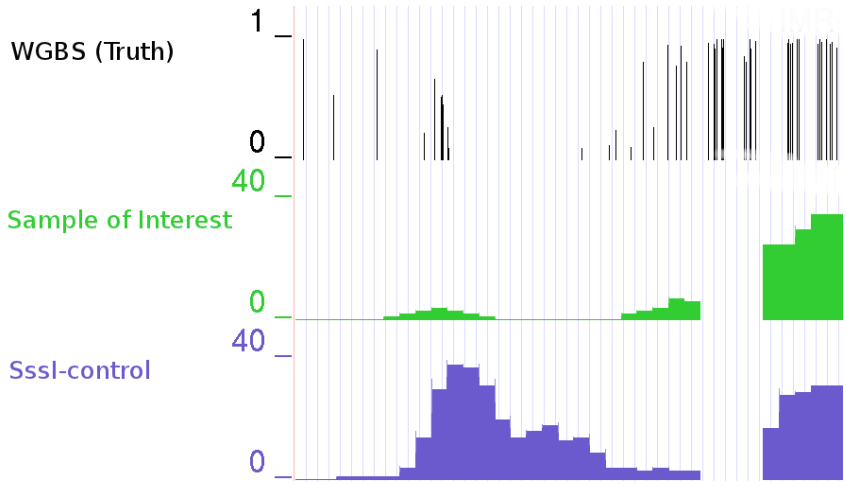
1. can distinguish inefficient capture from low methylation,
2. gives variance estimates,
3. accounts for copy-number-variations,
4. is computationally light,
5. is integrated into public domain and open source software (e.g. Bioconductor) to be directly applicable to routine tasks.



## Idea: Borrow strength from control information

Use an artificially full methylated (SssI-treated) control sample

1. to learn where the immunoprecipitation assay works.
2. to interpret the read density.






## BayMeth: Model formulation

Riebler et al., 2012, Tech Rep

We consider genomic regions $i = 1, \ldots, n$ and define

- $y_{i,C}$: Number of reads for the fully methylated (SssI) control.
- $y_{i,S}$: Number of reads for the sample of interest.

$$y_{i,C}|\lambda_i \sim \text{Poisson}(\lambda_i); \qquad y_{i,S}|\lambda_i, \mu_i \sim \text{Poisson}(f \times \lambda_i \times \mu_i)$$

with

$\lambda_i$: region-specific read density

$\mu_i$: the regional methylation level (Main parameter of interest)

f: known relative offset.

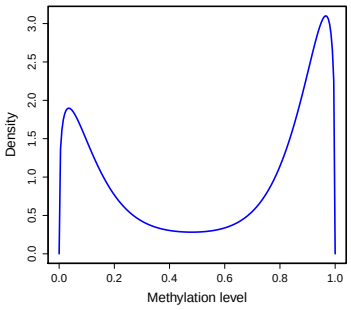
# Model formulation (II): Prior distributions

In a Bayesian framework, prior distributions are assigned to all parameters.

- For $\mu_i$: a (mixture of) beta distributions:

$$\mu_i \sim \sum_{m=1}^{M} w_m \, \text{Beta}(a_m, b_m),$$

with $0 \leq w_m \leq 1$, and $\sum_{m=1}^{M} w_m = 1$.




(In the simplest case a uniform prior: $M = 1$, $a_m = b_m = 1$).

- For $\lambda_i$: a gamma distribution with shape $\alpha$, rate $\beta$.



# Closed-form posterior marginal distribution

Notably, the marginal posterior distribution of the methylation level:

$$\text{p}(\mu_i | y_{i,s}, y_{i,c}) = \int_0^\infty \overbrace{\text{p}(\lambda_i, \mu_i | y_{i,s}, y_{i,c})}^{\text{Posterior distribution}} d\lambda_i$$

$$\overset{\text{cond.indep}}{=} \int_0^\infty \frac{\text{p}(\lambda_i)\,\text{p}(\mu_i)\,\text{p}(y_{i,c}|\lambda_i)\,\text{p}(y_{i,s}|\lambda_i, \mu_i)}{\text{p}(y_{i,s}, y_{i,c})} d\lambda_i.$$

is available in closed form.

Summary estimates:

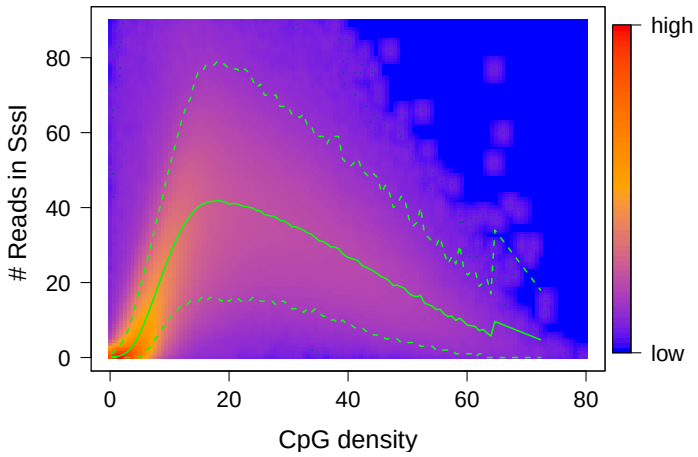
- Posterior mean and variance are analytically available and efficient to compute.
- Credible intervals can be computed numerically.

## Find prior parameters using empirical Bayes (EB)

1. Specify prior format for $\mu_i$ (i.e. number of beta components).
2. Divide regions into groups based on:
   - CpG density.
   - Sequence context (promoter, gene body, rest).
3. Determine parameters using EB for each group (in parallel).

Observations:

- context-specific information lead to biased results.
- uniform prior (Beta(1, 1)) on $\mu_i$ outperforms weighted beta mixtures.






## Software: Integration into `Repitools`-package

- Implementation in `R`.
- S4 class system.
- Computationally demanding tasks are done in `C`.
- Parallelisation over bins using the R-package `snowfall`.
- Integration into the Bioconductor R-package `Repitools` is in progress, so that it is soon available for routine tasks.

# Data flow (in progress)

```
> showClass("BayMethList")

Class "BayMethList" [package "Repitools"]

Slots:
Name:         windows         control sampleInterest         cpgDens
Class:        GRanges          matrix          matrix         numeric

Name:               f       priorTab        methEst
Class:         matrix           list           list
> bm <- BayMethList(windows=windows, control=co, sampleInterest=sI, cpgDens=cpgdens)
> ## Estimate the normalising offset f based on an MA-plot.
> bm <- determineOffset(bm, controlPlot=list(show=FALSE,
+          nsamp=50000, mfrow=c(1,1), ask=FALSE))

> ## Derive prior parameters using EB for "ngroups" CpG density classes.
> ## Use a mixture with "ncomp" components for the methylation level.
> bm <-  empBayes(bm, ngroups=100, ncomp=1, ncpu=NULL)

> ## Get mean and variance estimates and potentially credible intervals.
> bm <- methylEst(bm, ncomp=1, controlCI=list(compute=FALSE, method="quantile",
+         level=0.95, ncpu=NULL, ...))
```

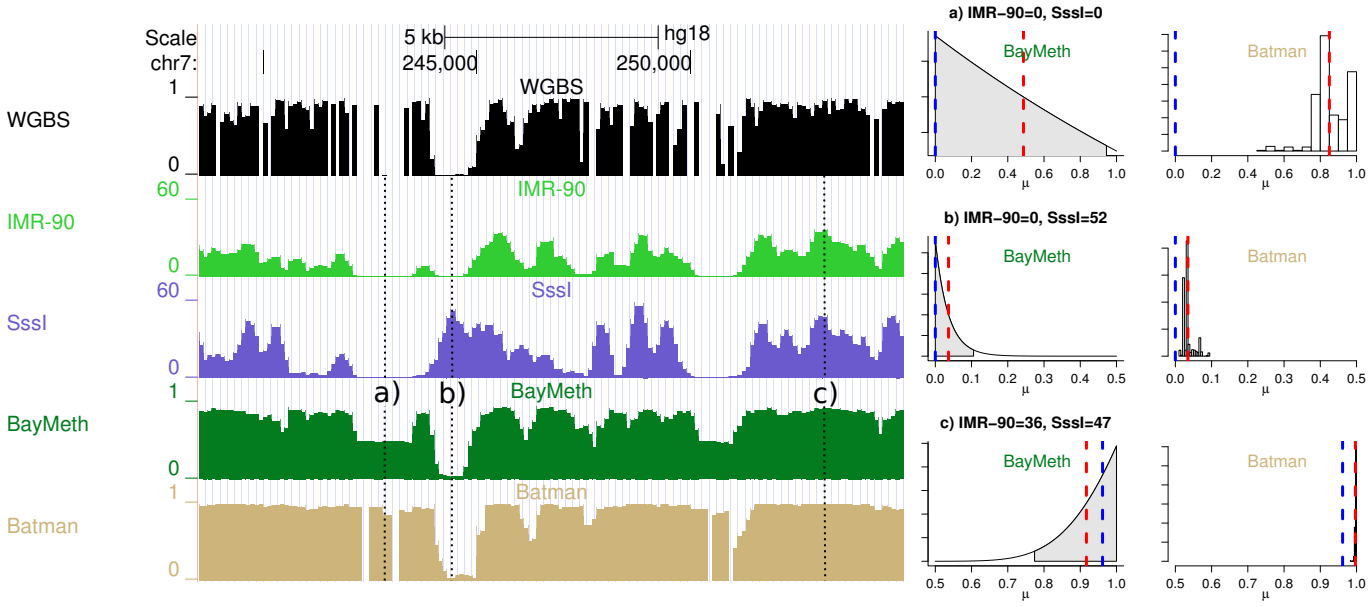
Mean and variance derivation in a genome-wide analysis $\approx$ 3 min.



# Applications: Lung fibroblast cell line (IMR-90)

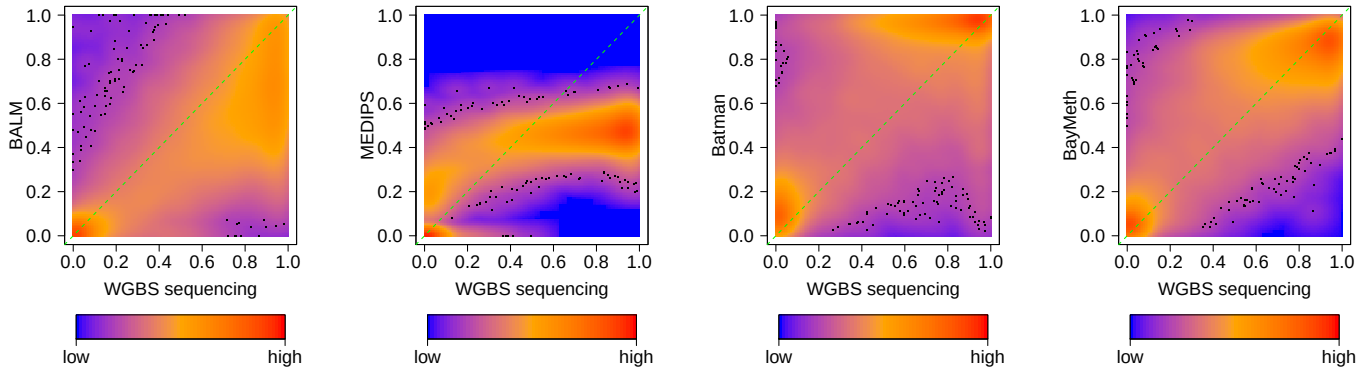
- True "true methylation estimates" of WGBS are available.

  Lister et al., 2009, Nature



# Performance assessment

- Scatter plots



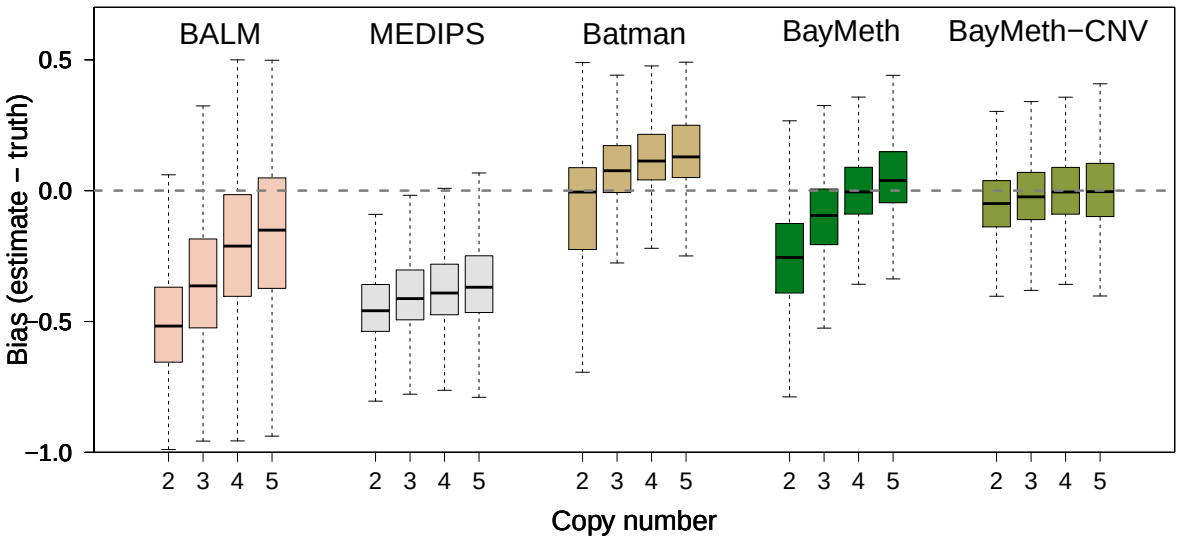

- Best performance in terms of:
  - correlation,
  - bias,
  - coverage probabilities.



# Prostate cancer cell line (genomewide)

Let $\text{cn}_i$ be the regional copy number state and ccn the most prominent state:

$$y_{i,s}|\mu_i, \lambda_i \sim \text{Poisson}(f \times \text{cn}_i \,/\, \text{ccn} \times \mu_i \times \lambda_i);$$



## Summary and Discussion

- Presentation of a novel Bayesian approach for affinity-capture-based DNA methylation analysis, which
  - leads to analytical expressions for the mean and variance.
  - provides credible intervals.
  - allows us to explicitly model copy number variation.
  - is user-friendly and computationally efficient.
- Broad utility of the method due to need of SssI control?
  - Better outcome compensates for a bit more work/money.
  - Making SssI control data available that others can utilise.



## Acknowledgments


- Mark Robinson
- Sue Clark, Jenny Song, Aaron Statham, Clare Stirzaker, Mirco Menigatti, Nadiya Mahmud, Charles Mein.
- Financial support by the “URPP-Grant” and the “Forschungskredit” for young researchers of the University of Zurich.


**Thank you for your attention!**